Speech recognition system improvements for North Sámi Speaker-dependent and Speaker Independent Tasks

Aku Rouhe, Mikko Kurimo

Department of lnformation and Communications Engineering
Aalto University, Finland

aku.rouhe@aalto.fi

Abstract

We are working on North Sámi, an under-resourced language, for which we have less than ten hours of transcribed speech in total. Previously, we applied wav2vec 2.0 pretrained large Transformer models to this data. However, error rates were still high. Here, we present a series of system improvements to these models, yielding minor performance improvements. We also experiment with a slightly larger text corpus, which provides a further minor performance improvement. Nonetheless, we conclude that more transcribed speech is needed, at least so that standard size development and test sets can be created.

Index Terms: Speech Recognition, North Sámi, SSL

1. Introduction

In this work, we present our work with optimizing speech recognition systems for North Sámi. We experiment with system improvements, which are motivated by the under-resourced scenario. Both on the North Sámi data, as well as on a simulated under-resourced scenario on Finnish, we find our designed hyperparameter choices yield minor improvements. Additionally, we find that some additional North Sámi text data also yields marginal improvements. In the end, we conclude that more transcribed speech in North Sámi is needed for substantial improvements to our systems.

In North Sámi we are working with very limited resources. Because there are no official test sets, it is challenging to split the data into training, validation, and test subsets of meaningful size. The data is also limited in the number of speakers, so creating and testing speaker-independent speech recognisers is an additional difficulty.

Previously, we applied wav2vec 2.0 pretrained Transformers to North Sámi [1]. We found that the hidden Markov model / deep neural network (HMM/DNN) approach outperformed the attention-based encoder-decoder (AED) approach, and thus are continuing our work here focusing on HMM/DNN-systems.

2. Related work

North Sámi (*Davvisámegiella*) is the largest of the Sámi languages by both number of speakers and by the land area in which it is spoken. It is spoken by approximately 20 000 to 25 000 people, in areas which fall inside Norway, Sweden, and Finland. North Sámi belongs to the Uralic language family. For spoken-language technology, it is an under-resourced, but actively researched language, and for instance a spell-checker has been created [2], as well as experiments in speech recognition and speech synthesis, enabling interactive technology [3].

Recently in speech recognition, large Self-Supervised Learning (SSL) pretrained Transformer models have achieved incredible performance with very limited resources, with learning criteria such as wav2vec 2.0 [4], HuBERT [5], and wavLM [6]. As little as ten minutes of transcribed speech has been enough to provide single digit WER on Librispeech (a relatively easy English read speech task). The SSL pretrained Transformer-models have also yielded very promising results for under-resourced languages through multi- and cross-lingual approaches [7].

3. Data

We have access to two corpora: Giellagas North [8] and the *UIT-SME TTS Corpus*. Giellagas North has been gathered through interviews, which have been transcribed. The interviewees use three distinct dialects: Torne Sámi, Finnmark Sámi and Sea Sámi, and the interviewers' Sámi speech is also included in the corpus. The interviewers may speak North Sámi as a second-language, and some interviewers use Finnish, which we filter out. Giellagas North contributes 19 speakers, totalling just two hours of material. The UIT-SME TTS Corpus has about eight hours of clean, prepared speech from two speakers.

We split Giellagas North by the annotated utterance segment boundaries. Since the UIT-SME TTS Corpus is distributed as long recordings, we use a preliminary hidden Markov model / Gaussian Mixture Model (HMM/GMM) system to segment the corpus into short utterances, divided at sentence-ending punctuation. We remove punctuation and capitalization from both corpora.

In our earlier work, we used the speech transcripts for language models. This decision allowed direct comparisons between End-to-End AED-models (which only use speech and transcripts) and HMM/DNN-systems. Here, we are able to leverage an additional North Sámi text resource. This resource, called Freecorpus (FC), consists of freely available texts, collected by Giellatekno and Divvun[1].

In addition to the North Sámi data, we present results on a simulated under-resourced scenario in Finnish. The Finnish data is a subset of the Finnish Parliament ASR Corpus [9].

3.1. Data splits

We use data splits from our earlier work [1].

For North Sámi we take two splits: one speaker-independent task, and one speaker-dependent task. In the speaker-dependent task, the same speakers appear in the training, validation, and test sets, while in the speaker-independent task, each data subset has different speakers (the validation speakers are also distinct from the test speakers). Both tasks

[1] https://giellalt.github.io/ling/corpus_repositories.html

use the UIT-SME TTS Corpus in the training data, and split Giellagas North in across the training, validation, and test subsets.

In Finnish, we use the Many-Speakers split, a speaker-independent task. It has many speakers in the training data, which helps in learning speaker-independence, although there is only a small amount of data per speaker.

Table 1 lists the three tasks' data details.

4. Speech recognition Systems

We base our systems on the HMM-system recipe developed in [1]. This section presents the existing recipe.

The recipe provides an HMM/GMM, which is used for forced time-alignment of the training data. The HMM/GMM system follows the standard Kaldi-toolkit [10] steps: (1) a simple monophone model, (2), a triphone model, (3) a triphone model with spliced input features and Maximum Likelihood Linear Transform (MLLT), (4) a triphone model with spliced input features, MLLT, and a feature-space Maximum Likelihood Linear Regression adaptation. The alignments from the HMM/GMM are used both as Cross-Entropy label targets as well as in the acoustic model state-tying algorithm. This was found to be helpful, despite the HMM/GMM completely failing on the test data (99% error rate). An initial HMM/GMM (not part of the final recipe) is also used to segment the UIT-SME TTS Corpus by punctuation, because the UIT-SME Corpus is distributed as long recordings.

As the final step of the recipe, a wav2vec 2.0 -based acoustic model is finetuned for the North Sámi data. Pretrained models with the Large architecture (approximately 300 million parameters) are used. Specifically, the Uralic V2 model[2] is used. Two fully-connected layers, with ReLU activations, Dropout, and BatchNorm, are added on top of the wav2vec 2.0, and the model has two output heads. The outputs are trained in a multi-task setting: one with Cross-Entropy, the other with the Lattice-Free Maximum Mutual Information (LF-MMI) criterion. The two output heads are also used at decode time, linearly combining the log-likelihoods (after log-Softmax) from each output. The acoustic models are implemented using a mix of SpeechBrain [11] and Kaldi. The acoustic models use grapheme-based lexica, which means that any difference in pronunciation across the three North Sámi dialects has to be learned by the model implicitly.

The language models use 400 subword units (SentencePiece Byte Pair Encoding[12]). They are long-span (10-gram) modified Kneser-Ney backoff models trained on the transcripts of the acoustic model training data using the variKN-toolkit[13].

We refer the reader to the original work for implementation details and hyperparameters.

5. Experiments and results

The recipe from original work, as described in Section 4, serves as the baseline. We set up a series of cumulative changes to the baseline, and test those resulting systems on the speaker-independent and speaker-dependent tasks. Our implementations are available online[3].

The North Sámi data splits leave very little data into the validation and test sets, which may impact the statistical evidence that experiments provide. To quantify this issue, we use a bootstrapped confidence-estimate [14]. This procedure gives an estimate of the probability that one system improves over another. We will call an estimated probability larger than 95% a *significant* improvement.

5.1. System A

In System A, we use an output normalization loss on the LF-MMI output head. This discourages the model to use extreme output values, which might result from overfitting to the training data. Additionally, we normalize the Cross-Entropy output by an empirical prior. This was found to yield consistently better results [15].

5.2. System B

In System B, we add SpecAugment [16] to System A. SpecAugment is a simple, but very effective data augmentation method, which should allow the model to avoid overfitting on the training data and help to learn noise-robust representations.

5.3. System C

In System C, we add label smoothing (smoothing value 0.1) to System B's cross-entropy loss. Label smoothing punishes overconfident predictions.

5.4. System XLSR

In System XLSR, we change System C to a different wav2vec 2.0 pretrained model. Instead of the Uralic V2 model, we use the XLS-R model[4], which has been pretrained on over 400 thousand hours of speech in 128 languages. The model is the same size as the Uralic V2 one.

5.5. System XLSR + FC-LM

Finally, in System XLSR + FC-LM, we add the North Sámi Freecorpus and train a new language model. This is only applied to the North Sámi data, naturally.

5.6. Some abandoned ideas

We report some ideas we tried, but which did not appear fruitful in preliminary experiments. We tried using just the first 15 Transformer layers of the wav2vec 2.0 Large architecture (sometimes reported to be helpful), but this yielded approximately similar performance. We tried decreasing the number of acoustic states, with the idea that less unit granularity would decrease chances of overfitting. This did not improve results. We tried training the output layers for some thousands of updates before switching to training the whole network, with the idea that randomly initialized output layers may lead to large gradients in the wav2vec 2.0 layers, which in turn might lead to catastrophic forgetting. However, our models do not seem to suffer from catastrophic forgetting, and this idea also did not yield improvements. Finally, we tried decreasing the size of the language model vocabulary, to lessen language model data sparsity, but this only yielded worse results.

[2] https://huggingface.co/facebook/wav2vec2-large-uralic-voxpopuli-v2

[3] https://github.com/aalto-speech/kaldi-sb-north-sme

[4] https://huggingface.co/facebook/wav2vec2-xls-r-300m

Table 1: *The data splits used in this work.*

	Number of Speakers	Speaker overlap	Number of Utterances	Size [hours]
All North Sámi Speaker Independent				
Train	7		5545	8.01
Validation	4		287	0.16
Test	10		1869	1.51
All North Sámi Speaker Dependent				
Train	21	✓	6960	9.14
Validation	11		110	0.08
Test	11		631	0.48
Finnish Parliament Many-Speakers				
Train	340		6668	20.09
Validation	10		954	2.76
Test	10		962	2.81

5.7. Results

Table 2 lists the results on the various data splits in this study. A bootstrap estimate confirms that the improvements on *All North Sámi Speaker Dependent* are significant for all systems ($>$ 99.99%). On *All North Sámi Speaker Dependent*, the improvements are significant for Systems C ($>$ 96.5%), System XLSR ($>$ 99.8%), and System XLSR + FC-LM ($>$ 99.99%), but for systems A and B, no statistically significant difference with the Baseline could be found.

Table 2: *Main results*

	WER/CER [%]	
	Validation	Test
All North Sámi Speaker Independent		
Baseline	79.20 / 56.76	72.66 / 46.53
System A	78.54 / 49.84	71.36 / 40.00
System B	80.22 / 52.36	71.15 / 41.50
System C	77.80 / 51.40	70.52 / 41.93
System XLSR	77.52 / 49.37	69.09 / 38.93
System XLSR + FC-LM	**76.21 / 48.90**	**66.43 / 38.63**
All North Sámi Speaker Dependent		
Baseline	45.85 / 25.21	51.78 / 29.65
System A	47.80 / 22.27	52.02 / 25.99
System B	50.73 / 22.18	51.95 / 24.25
System C	47.48 / 22.08	50.85 / 24.71
System XLSR	**45.69** / 21.82	50.51 / 24.76
System XLSR + FC-LM	46.18 / **20.91**	**47.85 / 22.54**
Finnish Parliament Many-Speakers		
Baseline	13.93 / 4.93	10.13 / 3.34
System A	13.43 / 4.56	10.02 / 3.14
System B	13.29 / **4.45**	**9.65** / 3.05
System C	**13.14** / 4.47	9.66 / **3.04**
System XLSR	18.48 / 5.25	15.09 / 3.72

6. Discussion

On North Sámi, we see marginal improvements from the hyperparameter choices. Additional, but minor improvements are yielded by using the Freecorpus language model. It appears that for truly substantial improvements in our systems, more transcribed speech is needed. Mirroring our earlier results, we see that the Speaker Independent task remains much more difficult than the speaker dependent one.

On the Finnish data, we mostly see similar performance improvements from the new systems, as with North Sámi. However, the XLS-R wav2vec 2.0 pretrained weights result in a major degradation in performance in Finnish, whereas on North Sámi it yields the best results. This may be a result of the domain of Uralic speech in the Uralic V2 pretraining data. The Uralic pretraining data comes from the European Parliament sessions, which are probably similar in style as the Finnish Parliament data. On the other hand, this result may also indicate that for a language that is not directly included in the pretraining data, a larger amount of languages and data is better than keeping to a particular language family.

It would be possible to increase the amount of data and speaker variety in the training data while maintaining a meaningful test set through the use of leave-one-out cross-validation. However, that comes at the cost of orders-of-magnitude more computation. For hyperparameter studies, it may be more fruitful to use traditional, constant data splits.

Since North Sámi suffers from lack of official benchmarks, we had to use very small test sets. However, through the use of a bootstrap significance estimate, we were able to verify that our results are statistically significant.

7. Conclusions

We proposed some improvements to our wav2vec 2.0 -based speech recognition systems for North Sámi. In experiments, we validated that the system changes yield minor improvements on North Sámi tasks, as well as a simulated under-resourced Finnish task. We experimented with an additional text-only corpus of North Sámi, which also yielded minor improvements.

Experiments have shown that large self-supervised pretrained models can reach impressive results even with very limited data. Nevertheless, we concluded that only minor improvements are available through training improvements and more language model data. Something is needed: more data, or better in-built structure. Leveraging the linguistic understanding of a language is a common way to make up for lack of data in under-resourced scenarios. However, we believe acquiring more transcribed speech is also necessary, because it would allow the curation of standard size (e.g. five hours) development- and test sets on which results would hold more statistical significance. As evidenced by the Giellagas corpus, care needs to be taken to cover different dialects and speakers of North Sámi.

8. References

[1] A. Rouhe, A. Virkkunen, J. Leinonen, and M. Kurimo, "Low Resource Comparison of Attention-based and Hybrid ASR Exploiting wav2vec 2.0," in *Proc. Interspeech 2022*, 2022, pp. 3543–3547.

[2] L. Wiechetek, S. N. Moshagen, B. Gaup, and T. Omma, "Many shades of grammar checking – launching a constraint grammar tool for north sámi," in *Proceedigns of the NoDaLiDa 2019 Workshop on Constraint Grammar - Methods, Tools and Applications*, 2019.

[3] K. Jokinen, K. Hiovain, N. Laxström, I. Rauhala, and G. Wilcock, *DigiSami and Digital Natives: Interaction Technology for the North Sami Language*. Singapore: Springer Singapore, 2017, pp. 3–19. [Online]. Available: https://doi.org/10.1007/978-981-10-2585-3_1

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[5] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[6] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[7] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Proc. Interspeech 2022*, 2022, pp. 2278–2282.

[8] "Pohjoissaamen näytekorpus." [Online]. Available: http://urn.fi/urn:nbn:fi:lb-201407302

[9] A. Virkkunen, A. Rouhe, N. Phan, and M. Kurimo, "Finnish Parliament ASR Corpus," *Language Resources and Evaluation*, 2023.

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[11] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," 2021.

[12] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. [Online]. Available: https://aclanthology.org/D18-2012

[13] V. Siivola, T. Hirsimaki, and S. Virpioja, "On growing and pruning kneser–ney smoothed n-gram models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1617–1624, 2007.

[14] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in asr performance evaluation," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2004, pp. I–409.

[15] A. Rouhe, T. Grósz, and M. Kurimo, "Finnish Parliament ASR Corpus," *Submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023 in review.

[16] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.